published in P. Sevestre and L. Matyas (eds.),
The Econometrics of Panel Data, 3nd edition,
Chapter 7, 2008, 215-245

Chapter 1

PARAMETRIC BINARY CHOICE MODELS

by Michael Lechner, Stefan Lollivier and Thierry Magnac

Revision 1.3: December 1, 2005.

Contents

1. Introduction

Binary dependent data are a common feature in many areas of empirical economics as, for example, in transportation choice, the analysis of unemployment, labour supply, schooling decisions, fertility decisions, innovation behaviour of firms, etc. As panel data is increasingly available, the demand for panel data models coping with binary dependent variables is also increasing. Also, dramatic increases in computer capacity have greatly enhanced our ability to estimate a new generation of models. The second volume of this handbook contains several applications based on this type of dependent variable and we will therefore limit this chapter to the exposition of econometric models and methods.

There is a long history of binary choice models applied to panel data which can for example be found in Arellano and Honore (2001), Baltagi (2000), Hsiao (1992, 1995, 2003), Lee (2002) or Sevestre (2002) as well as in chapters of econometrics textbooks as for instance Greene (2003) or Wooldridge (2000). Some of these books and chapters do not devote much space to the binary choice model. Here, in view of other chapters in this handbook that address related nonlinear models (qualitative, truncated or censored variables, nonparametric models, etc.) we focus on the parametric binary choice model and some of its semiparametric extensions. The binary choice model provides a convenient benchmark case, from which many results can be generalised to limited dependent variable models such as multinomial discrete choices (Train, 2002), transition models in continuous time (Kamionka, 1998) or to structural dynamic discrete choice models that are not studied here.

We tried to be more comprehensive than the papers and chapters mentioned and we provide an introduction into the many issues that arise in such models. We also try not only to provide an overview of different models and estimators but also to make sure that the technical level of this chapter is such that it can easily be understood by the applied econometrician. For all technical details, the reader is referred to the specific papers.

Before we discuss different versions of the binary choice panel data models, define first the notation for the data generating process underlying the prototypical binary choice panel model:

$$y_{it} = \mathbf{1}\{y_{it}^* > 0\} \text{ for any } i = 1, \ldots, N \text{ and } t = 1, \ldots, T,$$

where $\mathbf{1}\{.\}$ is the indicator of the event between bracket and where the latent dependent variables y_{it}^* are written as:

$$y_{it}^* = X_{it}\beta + \varepsilon_{it}$$

where β denotes a vector of parameters, X_{it} is a $1 \times K$ vector of explanatory variables and error terms ε_{it} stand for other unobserved variables. Stacking the T observations of individual i,

$$Y_i^* = X_i\beta + \varepsilon_i,$$

where $Y_i^* = (y_{i1}^*, ., y_{iT}^*)$ is the vector of latent variables, $X_i = (X_{i1}, ., X_{iT})$ is the $T \times K$ matrix of explanatory variables and $\varepsilon_i = (\varepsilon_{i1}, ., \varepsilon_{iT})$ is the $T \times 1$ vector of errors.

We focus on the estimation of parameter β and of parameters entering the distribution function of ε_{it}. We do not discuss assumptions under which such parameters can be used to compute other parameters, such as causal effects (Angrist, 2001). We also consider balanced panel data for ease of notation although the general case of unbalanced panel is generally not much more difficult if the data is missing at random (see chapter XXX).

As usual in econometrics we impose particular assumptions at the level of the latent model to generate the different versions of the observable model to be discussed in the sections of this chapter. These assumptions concern the correlation of the error terms over time as well as the correlation between the error terms and the explanatory variables. The properties for various conditional expectations of the observable binary dependent variable are then derived. We assume that the observations are obtained by independent draws in the population of statistical units 'i', also called individuals in this chapter. Working samples that we have in mind are much larger in dimension N than in dimension T and in most cases we consider asymptotics in N holding T fixed although we report on some recent work on large T approximations. Time effects can then be treated in a determistic way. In this chapter we frequently state our results for an important special case, the panel probit model where error terms ε_i are assumed to be normally distributed.

In Section 2 of this chapter we discuss different versions of the static random effects model when the explanatory variables are strictly exogenous. Depending on the autocorrelation structure of the errors different estimators are available and we detail their attractiveness in each situation by trading-off their efficiency and robustness with respect to misspecification. Section 3 considers the static model when a time invariant unobservable variable is correlated with the time varying explanatory variables. The non linearity of binary choice models makes it pretty hard to eliminate individual fixed effects in likelihood functions and moment conditions, because the usual 'differencing out trick' of the linear model does not work except in special cases. Imposing quite restrictive assumptions is the price to pay to estimate consistently parameters of interest. Finally, section 4 addresses the important issue of structural dynamics for fixed and random effects, in other words cases when the explanatory variables include lagged endogenous variables or are weakly exogenous only.

2. Random effects models under strict exogeneity

In this section we set up the simplest models and notations that will be used in the rest of the chapter. We consider in this chapter that random effects models defined as in Arellano and Honoré (2001) as models where errors in the latent model are independent of the explanatory variables.[1] This assumption does not hold with respect to the explanatory variables in the current period only but also in all past and future periods so that explanatory variables are also considered in this section to be strictly exogenous in the sense that:

$$F_{\varepsilon_t}(\varepsilon_{it}|X_i) = F_{\varepsilon_t}(\varepsilon_{it}), \tag{1.1}$$

where $F_{\varepsilon_t}(\varepsilon_{it})$ denotes the marginal distribution function of the error term in period t. When errors are not independent over time, it will also at times be useful to impose a stronger condition on the joint distribution of the T errors terms over time, denoted $F_{\varepsilon}^{(T)}(\cdot)$:

$$F_{\varepsilon}^{(T)}(\varepsilon_i|X_i) = F_{\varepsilon}^{(T)}(\varepsilon_i). \tag{1.2}$$

Note that as in binary choice models in cross-sections, marginal choice probabilities can be expressed in terms of the parameters of the latent model:

$$\begin{aligned} P(y_{it} = 1|X_i) &= E(y_{it} = 1|X_i) \\ &= E(y_{it} = 1|X_{it} = x_{it}) = 1 - F_{\varepsilon_t}(-X_{it}\beta). \end{aligned} \tag{1.3}$$

It also emphasizes that the expectation of a Bernoulli variable completely describes its distribution.

We already said that we would consider random samples only. Individual observations are independent and if θ generically denote all unknown parameters including those of the distribution function of errors, the sample likelihood function is the product of individual likelihood functions:

$$L(\theta) = \prod_{i=1}^{N} L_i(Y_i|X_i;\theta)$$

where $Y_i = (y_{i1}, ., y_{iT})$ is the vector of binary observations.

2.1 Errors are independent over time

When errors are independent over time, the panel model collapses to a cross-sectional model with NT independent observations and the maximum likelihood estimator is the standard estimator of choice. The likelihood function for one observation is given by:

$$L_i(Y_i|X_i;\theta) = \prod_{t=1}^{T}[1 - F_{\varepsilon_t}(-X_{it}\beta)]^{y_{it}} F_{\varepsilon_t}(-X_{it}\beta)^{(1-y_{it})}. \tag{1.4}$$

Later it will be pointed out that even if true errors are not independent over time, nevertheless the pseudo-maximum likelihood estimator (incorrectly) based on independence – the so called 'pooled estimator' - has attractive properties (Robinson, 1982).

Let $\Phi(\cdot)$ denote the cumulative distribution function (cdf) of the univariate zero mean unit variance normal distribution, we obtain the following log-likelihood function for the probit model :

$$L_i(Y_i|X_i;\beta,\sigma_2,...,\sigma_T;\sigma_1 = 1) = \\ \sum_{t=1}^{T} y_{it} \ln \Phi(\frac{X_{it}\beta}{\sigma_t}) + (1 - y_{it}) \ln[1 - \Phi(\frac{X_{it}\beta}{\sigma_t})].$$

Note that to identify the scale of the parameters, the standard error of the error term in the first period is normalised to 1 ($\sigma_1 = 1$). If all coefficients are allowed to vary over time in an unrestricted way, then more variances have to be normalised.[2] In many applications however, the variance of the error is kept constant over time ($\sigma_t = 1$). For notational convenience this assumption will be maintained in the remainder of the chapter.

2.2 One factor error terms

2.2.1 The model.

Probably the most immediate generalisation of the assumption of independent errors over time is a one-factor structure where all error terms are decomposed into two different independent components. One is constant over time (u_i) and is called the individual effect, the other one being time variable (v_{it}), but identically and independently distributed (iid) over time and individuals. Thus, we assume that for $i = 1, \ldots, N$ and $t = 1, \ldots, T$:

$$\varepsilon_{it} = u_i + v_{it}, \quad F_v^{(T)}(v_{i1}, ..., v_{iT}|X_i) = \prod_{t=1}^{T} F_{v_t}(v_{it});$$

$$F_{u,v}^{(T)}(u_i, v_{i1}, ..., v_{iT}|X_i) = F_u(u_i) \prod_{t=1}^{T} F_{v_t}(v_{it}).$$

The individual effect, u_i, can be interpreted as describing the influence of time-independent variables which are omitted from the model and that are independent of the explanatory variables. Note that the one-factor decomposition is quite strong in terms of its time series properties, because the correlation between the error terms of the latent model does not die out when the time distance between them is increased.

To achieve identification, restrictions need to be imposed on the variances of each error component which are denoted σ_v^2 and σ_u^2. For example, variance σ_v^2 can be assumed to be equal to a given value (to 1 in the normal case), or one can consider the restriction that the variance of the sum of error terms is equal to 1 ($\sigma_u^2 + \sigma_v^2 = 1$). It simplifies the comparison with cross section estimations. In this section, we do not restrict σ_u and σ_v for ease of notation though such a restriction should be imposed at the estimation stage.

2.2.2 Maximum likelihood estimation. The computation of the log-likelihood function is difficult when errors are not independent over time or have not a one-factor structure since the individual likelihood contribution is defined as an integral with respect to a T dimensional distribution function. Assumptions of independence or one-factor structure simplify the computation of the likelihood function (Butler and Moffit 1982).

The idea is the following. For a given value of u_i, the model is a standard binary choice model as the remaining error terms v_{it} are independent between dates and individuals. Conditional on u_i, the likelihood function of individual i is thus:

$$L_i(Y_i|X_i, u_i; \theta) = \prod_{t=1}^{T} \left[[1 - F_v(-X_{it}\beta - u_i)]^{y_{it}} [F_v(-X_{it}\beta - u_i)]^{1-y_{it}} \right]$$

The unconditional likelihood function is derived by integration:

$$L_i(Y_i|X_i; \theta) = \textstyle\int_{-\infty}^{+\infty} L_i(Y_i|X_i, u_i; \theta) f_u(u_i) du_i. \tag{1.5}$$

The computation of the likelihood function thus requires simple integrations only. Moreover, different parametric distribution functions for

u_i and v_{it} can be specified in this 'integrating out' approach. For instance, the marginal distribution functions of the two error components can be different as in the case with a normal random effect and logistic iid random error.[3] Also note that the random effect may be modelled in a flexible way. For example Heckman and Singer (1984), Mroz (1999), and many others suggested the modeling framework where the support of individual effects of u_i is discrete so that the cumulative distribution function of u_i is a step function. Geweke and Keane (2001) also suggest mixtures of normal distribution functions.

For the special case of a T normal variate error, u_i, the log-likelihood of the resulting probit model is given by:

$$L_i(Y_i|X_i;\theta) = \tag{1.6}$$

$$= \int_{-\infty}^{+\infty} \left\{ \prod_{t=1}^{T} \left[\Phi(\frac{X_{it}\beta + \sigma_u u_i}{\sigma_v})]^{y_{it}} [1 - \Phi(\frac{X_{it}\beta + \sigma_u u_i}{\sigma_v})]^{1-y_{it}} \right] \right\} \phi(u_i) du_i,$$

where $\phi(\cdot)$ denotes the density function of the standard normal distribution. In this case, the most usual identification restriction is $\sigma_u^2 + \sigma_v^2 = 1$, so that the disturbances can be written as:

$$\varepsilon_{it} = \gamma u_i + \sqrt{1 - \gamma^2} v_{it},$$

where u_i and v_{it} are univariate normal, $N(0,1)$, and $\gamma > 0$. Parameter γ^2 is the share of the variance of the error term due to individual effects.

The computation of the likelihood function is a well-known problem in mathematics and is performed using gaussian quadrature. The most efficient method of computation that leads to the so called 'random effects probit estimator' uses the Hermite integration formula (Butler and Moffit, 1982). See also the paper by Guilkey and Murphy (1993) for more details on this model and estimator as well as Lee (2000) for more discussion about the numerical algorithm.

Finally, Robinson (1982) and Avery, Hansen and Hotz (1983) show that the pooled estimator is an alternative to the previous method. The pooled estimator is the pseudo-maximum likelihood estimator where it is incorrectly assumed that errors are independent over time. As a pseudo likelihood estimator, it is consistent though inefficient. Note that the standard errors of estimated parameters are to be computed using pseudo-likelihood theory (Gouriéroux, Monfort and Trognon, 1984).

2.3 General error structures

Obviously, the autocorrelation structure implied by the one factor-structure is very restrictive. Correlations do not depend on the distance between periods t and t'. The general model that uses only the restrictions implied by equations (1.1) and (1.2) poses, however severe computational problems. Computing the maximum likelihood estimator requires high dimensional numerical integration. For example, Gaussian quadrature methods for the normal model do not work in practice when the dimension of integration is larger than four.

There are two ways out of these computational problems. First, instead of computing the exact maximum likelihood estimator, we can use simulation methods and approximate the ML estimator by simulated maximum likelihood (SML). It retains asymptotic efficiency under some conditions that will be stated later on (e.g. Hajivassiliou, Mc Fadden, Ruud, 1996). In particular, SML methods require that the number of simulations tends to infinity to obtain consistent estimators. As an alternative there are estimators which are more robust to misspecifications in the serial correlation structure but which are inefficient because they are either based on misspecified likelihood functions (pseudo-likelihood) or on moment conditions that do not depend on the correlation structure of the error terms (GMM, e.g. Avery, Hansen and Hotz, 1983, Breitung and Lechner, 1997, Bertschek and Lechner, 1998, Inkmann, 2000). Concerning pseudo-ML estimation, we already noted that the pooled probit estimator is consistent irrespective of the error structure. Such a consistency proof is however not available for the one-factor random effects probit estimator.

Define the following set function :

$$D(Y_i) = \left\{ Y_i^* \in \mathbb{R}^T \text{ such that } \begin{array}{c} 0 \leq y_{it}^* < +\infty \text{ if } y_{it} = 1 \\ -\infty < y_{it}^* < 0 \text{ if } y_{it} = 0 \end{array} \right\} \tag{1.7}$$

The contribution of observation i to the likelihood is:

$$L_i(Y_i \,|X_i; \theta) = E\left[\mathbf{1}\left\{Y_i^* \in D(Y_i)\right\}\right] \tag{1.8}$$

In probit models, ε_i is distributed as multivariate normal $N(0, \Omega)$, Ω being a $T \times T$ variance-covariance matrix. The likelihood function is:

$$L_i(Y_i|X_i; \theta) = \int_{D(Y_i)} \phi^{(T)}(Y_i^* - X_i\beta, \Omega) dY_i^*,$$

where $\phi^{(T)}(\cdot)$ denotes the density of the T-variate normal distribution.

In the general case, the covariance matrix of the errors Ω is unrestricted (except for identification purposes, see above). It is very frequent however to restrict its structure to reduce the number of parameters to be estimated. The reason for doing so is computation time, stability of convergence, occurrence of local extrema and the difficulties to pin down (locally identify) the matrix of correlations when the sample size is not very large. In many applications the random effects model discussed in the previous section is generalised by allowing for an AR(1) process in the time variant error component (v_{it}). Other more general structures however are feasible as well if there are enough data.

We will see below how to use simulation to approximate the likelihood function by using Simulated Maximum Likelihood (SML). Another popular estimation method consist in using conditional moments directly. They are derived from the true likelihood function and are approximated by simulation (Method of Simulated Moments or MSM). McFadden (1989) proposed to consider all possible sequences of binary variables over T periods, Y_ω, where ω runs from 1 to 2^T. Choice indicators are defined as $d_{i\omega} = 1$ if i chooses sequence ω and is equal to 0 otherwise. A moment estimator solves the empirical counterpart of the moment condition:

$$E\left[\sum_{\omega=1}^{2^T} W_{i\omega}\left[d_{i\omega} - P_{i\omega}(\theta)\right]\right] = 0, \tag{1.9}$$

where $P_{i\omega}(\theta) = L_i(Y_\omega \mid X_i, \theta)$ is the probability of sequence ω (i.e. such that $Y_i = Y_\omega$). The optimal matrix of instruments $W_{i\omega}$ in the moment condition is:

$$W_{i\omega}^* = \left.\frac{\partial \log[P_{i\omega}(\theta)]}{\partial\theta}\right|_{\theta=\theta_0},$$

where parameter θ_0 is the true value of θ. In practice, any consistent estimator is a good choice to approximate parameter θ_0. The first of a two-step GMM procedure using the moment conditions above and identity weights can lead to such a consistent estimate. It is then plugged in the expression for $W_{i\omega}^*$ at the second step.

Even if T is moderately large however, the number of sequences ω is geometric in T (2^T) and functions $P_{i\omega}(\theta)$ can be very small. What proposes Keane (1994) is to replace in equation (1.9), unconditional probabilities by conditional probabilities:

$$E\left[\sum_{t=1}^{T}\sum_{j=0}^{1}\tilde{W}_{itj}\left(d_{itj}-P_{itj}(\theta)\right)\right]=0,$$

where $d_{itj}=1$ if and only if $y_{it}=j$ and where:

$$\begin{aligned}P_{itj}(\theta)&=P(y_{it}=j\mid y_{i1},.,y_{it-1},X_i;\theta)\\&=\frac{P(y_{it}=j,y_{i1},.,y_{it-1}\mid X_i;\theta)}{P(y_{i1},.,y_{it-1}\mid X_i;\theta)}\end{aligned}$$

is the conditional probability of choice j conditional on observed lagged choices.

Finally, maximising the expectation of the log-likelihood function $E\log[L_i(Y_i\mid X_i,\theta)]$ is equivalent to solving the following system of score equations with respect to θ:

$$E\left[S_i(\theta)\right]=0,$$

where $S_i(\theta)=\frac{\partial\log[L_i(Y_i|X_i,\theta)]}{\partial\theta}$ is the score function for individual i. It can be shown that, in most limited dependent variable models (Hajivassiliou and McFadden, 1998):

$$\frac{\partial}{\partial\theta}L_i(Y_i\mid X_i,\theta)=E\left[g_i(Y_i^*-X_i\beta)\mathbf{1}\left\{Y_i^*\in D(Y_i)\right\}\right]$$

where:

$$g_i(u)=\left[\begin{array}{c}X_i'\Omega^{-1}u\\\Omega^{-1}(uu'-\Omega)\Omega^{-1}/2\end{array}\right]$$

The score function can then be written as a conditional expectation:

$$S_i(\theta)=E\left[g_i(Y_i^*-X_i\beta)\left|Y_i^*\in D(Y_i)\right.\right]\tag{1.10}$$

which opens up the possibility of computing the scores by simulations (Method of Simulated Scores, MSS, Hajivassiliou and McFadden, 1998).

2.4 Simulation methods

Simulation methods (SML, MSM, MSS) based on the criteria established in the previous section consist in computing the expectation of a function of T random variates. The exact values of these high dimensional integrals are too difficult to compute and these expectations are approximated by sums of random draws using laws of large numbers:

$$\frac{1}{H}\sum_{h=1}^{H}f(\varepsilon_h)\underset{H\to\infty}{\overset{P}{\to}}Ef(\varepsilon)$$

when ε_h is a random draw from a distribution. In the case of panel probit models, it is a multivariate normal distribution function, $N(0, \Omega)$.

It is not the purpose of this chapter to review the general theory of simulation (see Gouriéroux and Monfort, 1996, Geweke and Keane, 2001). We review the properties of such methods in panel probit models only to which we add a brief explanation of Gibbs resampling methods which borrow their principle from Bayesian techniques.

2.4.1 The comparison between SML, MSM, MSS in probit models. The naive SML function is for instance:

$$\frac{1}{H} \sum_{h=1}^{H} I \left\{ Y_i^* \in D(Y_i) \right\}$$

where $I[Y_i^* \in D(Y_i)]$ is a simulator. It is not continuous with respect to the parameter of interest however and this simulation method is not recommendable. What is recommended is to use a smooth simulator which is differentiable with respect to the parameter of interest. The Monte Carlo evidence that the Geweke-Hajivassiliou-Keane (GHK) simulator is the best one in multivariate probit models seems overwhelming (see Geweke and Keane, 2001 and Hajivassiliou, McFadden, and Ruud, 1996, for a presentation).

The asymptotic conditions concerning the number of draws (H) and leading to consistency, absence of asymptotic bias and asymptotic normality are more or less restrictive according to each method, SML, MSM or MSS (Gouriéroux and Monfort, 1993). The method of simulated moments (MSM) yields consistent, asymptotically unbiased and normally distributed estimators as $N \to \infty$ when H is fixed because the moment condition (1.9) is linear in the simulated expression (or the expectation). In Keane's (1994) version of MSM where conditional probabilities are computed by taking ratios, the estimator is only consistent when the number of draws tends to infinity. Similarly, because a logarithmic transformation is taken, SML is not consistent when H is fixed. Consistency is obtained when H grows at any rate towards infinity (Lee, 1992). Furthermore, a sufficient condition to obtain asymptotically unbiased, asymptotically normal and efficient estimates is $\sqrt{N}/H \to 0$ as $N \to \infty$ (Lee, 1992, Gouriéroux and Monfort, 1993).

It is the reason why some authors prefer MSM to SML. As already said, MSM however requires the computation of the probabilities of all the potential paths with longitudinal data although the less intensive method proposed by Keane (1994) seems to work well in panel probit models (Geweke, Keane and Runkle, 1997). The computation becomes cumbersome when the number of periods is large and there is evidence that small sample biases in MSM are much larger than the simulation bias (Geweke and Keane, 2001). Lee (1995) proposed procedures to correct asymptotic biases though results are far from impressive (Lee, 1997, Magnac, 2000). The GHK simulator is an accurate simulator though it may require a large number of draws to be close to competitors such as Monte Carlo Markov Chains (MCMC) methods (Geweke, Keane and Runkle, 1997). There seems to be a general consensus between authors about the deterioration of all estimators when the amount of serial correlation increases.

Another way to obtain consistent estimators for fixed H is the method of simulated scores (MSS) if the simulator is unbiased. It seems that it is simpler than MSM because it implicitly solves the search for optimal instruments. Hajivassiliou and McFadden (1998) proposes an acceptance-rejection algorithm consisting in rejecting the draw if the condition in equation (1.10) is not verified. The simulator is not smooth however and as already said a smooth simulator seems to be a guarantee of stability and success for an estimation method. Moreover, in particular when T exceeds four or five, it is possible for some individuals that the acceptance condition is so strong that no draw is accepted. Other methods consist in considering algorithms either based on GHK simulations of the score or on Gibbs resampling. Formulas and an evaluation are given in Hajivassiliou, McFadden, and Ruud (1996).[4]

2.4.2 Gibbs sampling and data augmentation. It is possible however to avoid maximisation by applying Gibbs sampling techniques and data augmentation in multiperiod probit models (Geweke, Keane and Runkle, 1997, Chib and Greenberg, 1998, Chib, 2001). Though the original setting of Monte Carlo Markov Chains (MCMC) is Bayesian, it can be applied to classical settings as shown by Geweke, Keane and Runkle, (1997). The posterior density function of parameter θ given the data $(Y, X) = \{(Y_i, X_i); i = 1, ., n\}$ can indeed be used to compute posterior means and variance-covariance matrices to be used as classical estimators and their variance-covariance matrices.

To compute the posterior density $p(\theta \mid Y, X)$, we rely on two tools. One is the Metropolis-Hastings algorithm which allows for drawing samples in any (well behaved) multivariate density function, the other is Gibbs resampling which allows to draw in the conditional densities instead of the joint density function.

In the case of panel probit models, it runs as follows. First, let us “augment” the data by introducing the unknown latent variables $Y_i^* = X_i\beta + \varepsilon$ in order to draw from the posterior density $p(\theta, Y^* \mid Y, X)$ instead of the original density function. The reason is that it will be much easier to sample into density functions conditional on the missing latent variables. Second, parameter θ is decomposed into different blocks $(\theta_1, ., \theta_J)$ according to the different types of parameters in β or in Ω the variance-covariance matrix.[5]

Let choose some initial values for θ, say $\theta^{(0)}$ and proceed as follows. Draw Y^* in the distribution function $p(Y^* \mid \theta^{(0)}, Y, X)$ – it is a multivariate truncated normal density function – in a very similar way to the GHK simulator. Then draw a new value for the first block θ_1 in θ, i.e. from $p(\theta_1 \mid Y^*, \theta_{-1}^{(0)}, Y, X)$ where $\theta_{-1}^{(0)}$ is constructed from parameter $\theta^{(0)}$ by omitting $\theta_1^{(0)}$. Denote this draw $\theta_1^{(1)}$. Do similar steps for all blocks $j = 2, ., J$, using the updated parameters, until a new value $\theta^{(1)}$ is completed. Details of each step are given in Chib and Greenberg (1998). Repeat the whole step M times – M depends on the structure of the problem (Chib, 2001). Trim the beginning of the sample $\{\theta^{(0)}, ..., \theta^{(m)}\}$, the first 200 observations say. Then, the empirical density function of $\{\theta^{(m+1)}, ..., \theta^{(M)}\}$ is $p(\theta \mid Y, X)$. Once again, this method is computer intensive with large samples and many dates. It is however a close competitor to SML and MSS (Geweke and Keane, 2001).

2.4.3 Using marginal moments and GMM. Instead of working with the joint distribution function, the model defined by equation (1.8) implies the following moment conditions about the marginal period-by-period distribution functions.[6]

$$\begin{aligned}
&E[M(Y, X; \beta_0)|X] = 0,\\
&M(Y, X; \beta) = [m_1(y_1, X; \beta), ..., m_t(y_t, X; \beta), ..., m_T(y_T, X; \beta)]',\\
&m_t(y_t, X; \beta) = Y_t - [1 - F(-X_t\beta)].
\end{aligned} \tag{1.11}$$

For the probit model the last expression specialises to $m_t(Y_t, X_t; \beta) = y_t - \Phi(X_t\beta)$. Although the conditional moment estimator (CME) based on these marginal moments will be less efficient than full information maximum likelihood (FIML), these moment estimators have the clear advantage that fast and accurate approximation algorithms are available and that they do not depend on the off-diagonal elements of the covariance matrix of the error terms. Thus, these nuisance parameters need not be estimated to obtain consistent estimates of the scaled slope parameters of the latent model. At least, these estimators yields interesting initial conditions and previous methods can be used to increase efficiency.

As in the full information case, there remains the issue of specifying the instrument matrix. First, let us consider a way to use these marginal moments under our current set of assumptions in the asymptotically efficient way. Optimal instruments are given by:

$$A^*(X_i,\theta_0) = D(X_i,\theta_0)'\Omega(X_i,\theta_0)^{-1};$$

$$D(X_i,\theta) = E\frac{\partial M(Y,X_i,\theta)}{\partial\theta}|X = X_i; \tag{1.12}$$

$$\Omega(X_i,\theta) = E[M(Y,X_i,\theta)M(Y,X_i,\theta)']|X = X_i. \tag{1.13}$$

For the special case of the probit model under strict exogeneity the two other elements of (13) have the following form:

$$D_{it}(X_{it};\beta_0) = -\phi(X_{it}\beta_0)X_{it}$$

$$\omega_{its}(X_{it},\beta_0) = [E(Y_t - \Phi_{it})(Y_s - \Phi_{is})|X = X_i] \tag{1.14}$$

$$= \left\{ \begin{array}{l} \Phi_{it}(1-\Phi_{it}) \text{ if } t = s \\ \Phi_{its}^{(2)} - \Phi_{it}\Phi_{is} \text{ if } t \neq s \end{array} \right. \tag{1.15}$$

where $\Phi_{it} = \Phi(X_{it}\beta_0)$ and $\Phi_{its}^{(2)} = \Phi^{(2)}(X_{it}\beta_0, X_{is}\beta_0, \rho_{ts})$ denotes the cdf of the bivariate normal distribution with correlation coefficient ρ_{ts}. The estimation of the optimal instruments is cumbersome because they vary with the regressors in a nonlinear way and depend on the correlation coefficients.

There are several different ways to obtain consistent estimates of the optimal instruments. Bertschek and Lechner (1998) propose to estimate the conditional matrix nonparametrically. They focus on the k-nearest neighbour (k-NN) approach to estimate $\Omega(X_i)$, because of its simplicity. K-NN averages locally over functions of the data of those observations belonging to the k-nearest neighbours. Under regularity conditions (Newey, 1993), this gives consistent estimates of $\Omega(X_i)$ evaluated at $\tilde{\beta}_N$ and denoted by $\tilde{\Omega}(X_i)$ for each observation 'i' without the need for estimating ρ_{ts}. Thus, an element of $\Omega(X_i)$ is estimated by:

$$\tilde{\omega}_{its}(X_i) = \sum_{j=1}^{N} w_{ijts} m_t(y_{jt}, X_{jt}; \tilde{\beta}_N) m_s(y_{it}, X_{it}; \tilde{\beta}_N), \tag{1.16}$$

where w_{ijts} represents a weight function. This does not involve an integral over a bivariate distribution. For more details one different variants of the estimator and how to implement it, the reader is referred to Bertschek and Lechner (1998). In their Monte Carlo study it appeared that optimal (nonparametric) Conditional Moment estimators based on moments rescaled to have a homoscedastic variance performed much better in small samples. They are based on:

$$m_t^W(Y_t, X; \beta) = \frac{m_t(Y_t, X_t; \beta)}{\sqrt{E[m_t^{(}Y_t, X_t; \beta)^2 | X = X_i]}}. \tag{1.17}$$

The expression of the conditional covariance matrix of these moments and the conditional expectation of the first derivatives are somewhat different from the previous ones, but the same general estimation principles can be applied in this case as well.[7] Inkman (2000) proposes additional Monte Carlo experiments comparing GMM estimators to SML with and without heteroskedasticity.

2.4.4 Other estimators based on suboptimal instruments.

Of course there are many other specifications for the instrument matrix that lead to consistent, although not necessarily efficient, estimators for the slope coefficients. Their implementation as well as their efficiency ranking is discussed in detail in Bertschek and Lechner (1998). For example they show that the pooled probit estimator is asymptotically equivalent to the previous GMM estimator when the instruments are based on equations (1.16) to (1.13) but the off-diagonal elements of $\Omega(X_i)$ are set to zero. Avery, Hansen and Hotz (1983) also suggested to improve the efficiency of the pooled probit by exploiting strict exogeneity in another way by stacking the instrument matrix, so as to exploit that the conditional moment in period t is also uncorrelated with any function of regressors from other periods.

Chamberlain (1980) suggests yet another very simple route to improve the efficiency of the pooled probit estimator when there are arbitrary correlations of the errors over time which avoids setting up a 'complicated' GMM estimator. Since cross-section probits give consistent estimates of the coefficients for each period (scaled by the standard deviation of the period error term), the idea is to perform T probits period by period (leading to $T \times K$ coefficient estimates) and combine them in a second step using a Minimum Distance estimator. The variance-covariance matrix of estimators at different time periods should be computed to construct efficient estimates at the second step although small sample bias could also be a problem (Altonji and Segal, 1996). In the case of homoscedasticity over time this step will be simple GLS, otherwise a nonlinear optimisation in the parametric space is required.[8]

2.5 How to choose a random effects estimator for an application

This section introduced several estimators that are applicable in the case of random effect models under strict exogeneity. In practice the question is what correlation structure to impose and which estimator to use. Concerning the correlation structure, one has to bear in mind that exclusion restrictions are important for non parametric identification and thus that explanatory variables should be sufficiently variable across time in order to permit the identification of a very general pattern of correlation of errors. For empirical applications of the estimators that we have reviewed, the following issues seem to be important: Small sample performance, ease of computation, efficiency, robustness. We will address them in turn.

With respect to small sample performance of GMM estimators, Monte Carlo simulations by Breitung and Lechner (1997), Bertschek and Lechner (1998) and Inkmann (2000) suggest that estimators based on too many overidentifying restrictions (i.e. too many instruments), like the sequential estimators and some of the estimators suggested by Avery, Hansen, and Hotz (1983) are subject to typical weak instruments problem of GMM estimation due to too many instruments . Thus they are not very attractive for applications. The exactly identified estimators appear to work fine.

'Ease of computation' is partly a subjective judgement depending on computing skill and software available. Clearly, pooled probit is the easiest to implement, but random effects ML is available in many software packages as well. Exact ML is clearly not feasible for T larger than 4. For GMM and simulation methods, there is GAUSS code available on the Web (Geweke and Keane, 2001 for instance) but they are not part of any commercial software package. The issue of computation time is less important now that it was some time ago (Greene, 2002) and the simulation estimators are getting more and more implementable with the increase of computing power. Asymptotic efficiency is important when samples are large. Clearly, exact ML is the most efficient one and can in principle be almost exactly approximated by the simulation estimators discussed.

With respect to robustness, it is probably most important to consider violations of the assumption that explanatory variables at all periods are exogeneous and restrictions of the autocorrelation structure of the error terms. We will address the issue of exogeneity at the end of this chapter though the general conclusions are very close to the linear case, as far as we know. Concerning the autocorrelation of errors, pooled probit either in its pseudo-ML or GMM version is robust if it uses marginal conditional moments. It is not true for the other ML estimators that rely on the correct specification of the autocorrelation structure. Finally, GMM estimators as they have been proposed here (with the exception of pooled probit, of course) are robust against any autocorrelation. However, they obtain their efficiency gains by exploiting strict exogeneity and may become inconsistent if this assumption does not hold.

2.6 Correlated Effects

In the correlated effects (or unrelated effects) model, we abandon the assumption that individual effects and explanatory variables are independent. In analogy with the linear panel data case, Chamberlain (1984) proposes, in a random effect panel data nonlinear model, to replace the assumption that individual effects u_i are independent of the regressors by a weaker assumption. This assumption is derived from writing a linear regression:

$$u_i = X_i \gamma + \eta_i \tag{1.18}$$

where explanatory variables at all periods, X_i, are now independent of the redefined individual effect η_i. This parametrization is convenient but not totally consistent with the preceding assumptions : considering the individual effect as a function of the X_i variables makes its definition depend on the length of the panel. However, all results derived in the previous section can readily be applied by replacing explanatory variables X_{it} by the whole sequence X_i at each period[9].

To recover the parameters of interest, β, two procedures can be used. The first method uses minimum distance estimation and the so called $\pi-$matrix technique of Chamberlain (Crépon & Mairesse, 1995). The reduced form:

$$y_{it}^* = X_i \gamma_t + \eta_i + v_{it}, \tag{1.19}$$

is first estimated. The second step consists in imposing the constraints given by:

$$\gamma_t = \gamma + \beta e_t \tag{1.20}$$

where e_t is an appropriate known matrix derived from equations (1.18) and (1.19).

The second procedure uses constrained maximum likelihood estimation by imposing the previous constraint (1.20) on the parameters of the structural model.

The assumption of independence between η_i and X_i is quite strong in the non-linear case in stark contrast to the innocuous non-correlation assumption in the linear case. Moreover, it also introduces constraints on the data generating process of x_i if one wants to extend this framework when additional period information comes in (Honoré, 2002). Consider that we add a new period $T+1$ to the data and rewrite the projection as:

$$u_i = X_i\tilde{\gamma} + X_{iT+1}\tilde{\gamma}_{T+1} + \tilde{\eta}_i$$

By substracting both linear regressions at times T and $T+1$ and taking expectations conditional on information at period T, it implies that:

$$E(X_{iT+1} \mid X_i) = X_i(\gamma - \tilde{\gamma})/\tilde{\gamma}_{T+1}$$

which is not only linear in X_i but also, only depend on parameters governing the y_{it} process.

It is therefore tempting to relax equation (1.18) and admit that individual effects are a more general function of explanatory variables:

$$u_i = f(X_i) + \eta_i$$

where $f(.)$ is an unknown function satisfying weak restrictions (Newey, 1994). Even if the independence assumption between the individual effect η_i and explanatory variables x_i is still restrictive – because the variance of η_i is constant for instance – this framework is much more general than the previous one. What Newey (1994) proposes is based on the cross section estimation technique that we already talked about.

Consider the simple one-factor model where the variance of the individual-and-period specific shocks is not period-dependent, σ_v^2, and where the variance of η_i is such that $\sigma_v^2+\sigma_\eta^2$ is normalized to one. We therefore have:

$$E(y_{it} \mid X_i) = \Phi(X_{it}\beta + f(X_i))$$

where Φ is the distribution function of a zero-mean unit-variance normal variate. It translates into:

$$\Phi^{-1}(E(y_{it} \mid X_i)) = X_{it}\beta + f(X_i) \tag{1.21}$$

By any differencing operator (Arellano, 2003) and for instance by first differencing, we can eliminate the nuisance function $f(X_i)$ to get:

$$\Phi^{-1}(E(y_{it} \mid X_i)) - \Phi^{-1}(E(y_{it-1} \mid X_i)) = (X_{it} - X_{it-1})\beta \tag{1.22}$$

The estimation runs as follows. Estimates of $E(y_{it} \mid X_i)$ at any period are first obtained by series estimation (Newey, 1994) or any other non parametric method (kernel, local linear, smoothing spline, see Pagan & Ullah, 1999 for instance). A consistent estimate of β is then obtained by using the previous moment condition (1.22).

A few remarks are in order. First, Newey (1994) proposes such a modeling framework in order to show how to derive asymptotic variance-covariance matrices of semi-parametric estimators. As it is outside of the scope of this chapter, the reader is refered, for this topic, to the original paper. It can also be noted that as an estimate of $f(X_i)$ can be obtained, in a second step, by using the equation in levels (1.21). One can then use a random effect approach to estimate the serial correlation of the random vector, v_{it}. Finally, there is a non parametric version of this method (Chen, 1998) where Φ is replaced by an unknown function to be estimated, under some identification restrictions.

3. Fixed effects models under strict exogeneity

In the so-called fixed effect model, the error component structure of section 2.2 is assumed. The dependence between individual effects and explanatory variables is now unrestricted in contrast to the independence assumption in the random effects model. In this section, we retain the assumption of strict exogeneity that explanatory variables and period-and-individual shocks are independent. We write the model as:

$$y_{it} = \mathbf{1}\{\mathbf{X}_{it}\beta + u_i + v_{it} > 0\} \tag{1.23}$$

where additional assumptions are developed below.

As the conditional distribution of individual effects u_i is unrestricted, the vector of individual effects should be treated as a nuisance parameter that we should either consistently estimate or that we should eliminate. If we cannot eliminate the fixed effects, asymptotics in T are required in most cases.[10] It is because only T observations are available to estimate each individual effect. It cannot be consistent as $N \to \infty$ and its inconsistency generically contaminates the estimation of the parameter of interest. It gives rise to the problem of incidental parameters (Lancaster, 2000). The assumption that T is fixed seems to be a reasonable approximation with survey data since the number of periods over which individuals are observed is often small. At the end of the section however, we will see how better large T approximations can be constructed for moderate values of T.

The other route is to difference out the individual effects. It is more difficult in non-linear models than in linear ones because it is not possible to consider linear transforms of the latent variable and to calculate within-type estimators. In other words, it is much harder to find moment conditions and specific likelihood functions that depend on the slope coefficient but do not depend on the fixed effects. In short panels, ML or GMM estimation of fixed effects probit models where the individual effects are treated as parameters to be estimated are severely biased if T is small (Heckman, 1981a).

In the first sub- sections we discuss some methods that appeared in the literature that circumvent this problem and lead to consistent estimators for $N \to \infty$ and T is *fixed*. Of course, there is always a price to pay either in terms of additional assumptions needed or in terms of the statistical properties of these estimators.

3.1 The model

As already said, we consider equation (1.23) and we stick to the assumption of strict exogeneity of the explanatory variables:

$$F_{\varepsilon_t}(\varepsilon_{it}|u_i, X_{i1}, ..., X_{iT}) = F_{\varepsilon_t}(\varepsilon_{it}|u_i). \tag{1.24}$$

Using the error component structure of section 2.2, we can reformulate this assumption:

$$F_{v_t}(v_{it}|u_i, X_{i1}, ..., X_{iT}) = F_{v_t}(v_{it}). \tag{1.25}$$

Note that $F_{\varepsilon_t}(\varepsilon_{it}|X_{i1}, ..., X_{iT}) \neq F_{\varepsilon_t}(\varepsilon_{it})$ and also note that the distribution of the individual effect is unrestricted and can thus be correlated with observables. In most cases we will also impose that the errors are independent conditional on the fixed effect:

$$\begin{aligned} F(\varepsilon_{i1}, ..., \varepsilon_{iT}|u_i, X_{i1}, ..., X_{iT}) &= \prod_{t=1}^{T} F_{\varepsilon_t}(\varepsilon_{it}|u_i) \qquad (1.26) \\ F(v_{i1}, ..., v_{iT}|u_i, X_{i1}, ..., X_{iT}) &= \prod_{t=1}^{T} F_{v_t}(v_{it}). \end{aligned}$$

There are two obvious difficulties with respect to identification in such a model. First, it is impossible to identify the effects of time-invariant variables.[11] It has serious consequences because it implies that choice probabilities in the population are not identified. We cannot compare probabilities for different values of the explanatory variables. In other words, a fixed effect model that does not impose some assumption on distribution of the fixed effects cannot be used to identify causal (treatment) effects. This sometimes overlooked feature limits the use of fixed effects models.[12] What remains identified are the conditional treatment effects, conditional on any (unknown) value of the individual effect.

The second difficulty is specific to discrete data. In general, the individuals who stay all over the period of observation in a given state do not provide any information concerning the determination of the parameters. It stems from an identification problem, the so called mover-stayer problem. Consider someone which stays in state 1 from period 1 to T. Let v_i be any value of the individual-and-period shocks. Then if the individual effect u_i is a coherent value in model (1.23) with staying in the state all the time, then any value $\bar{u}_i \geq u_i$ is also coherent with model (1.23). Estimations are thus implemented on the sub-sample of people who move at least once between the two states ("moving" individuals).

3.2 The method of conditional likelihood

The existence of biases leads to avoid direct ML estimations when the number of dates is less than ten (Heckman, 1981a). In certain cases, the bias can consist in multiplying by two the value of some parameters (Andersen, 1971 ; Chamberlain, 1984 ; Hsiao, 1996). This features makes this estimator pretty unattractive in large N, small T type of applications. If the logit specification is assumed however, it is possible to set up a conditional likelihood function whose maximisation gives consistent estimators of the parameters of interest β, regardless the length of the time period.

Conditional logit: T periods. In the case where random errors, v_{it}, are independent over time and are logistically distributed, the sum $y_{i+} = \sum_{t=1}^{T} y_{it}$, is a sufficient statistic for the fixed effects, in the sense that the distribution of the data given y_{i+} does not depend on the fixed effect. Consider the logit model :

$$P(y_{it} = 1 | X_i, u_i) = F(X_{it}\beta + u_i), \tag{1.27}$$

where $F(z) = \frac{\exp(z)}{1+\exp(z)} = \frac{1}{1+\exp(-z)}$

The idea is to compute probabilities conditional on the number of times the individuals is in state 1:

$$L_i(\theta) = P(y_{i1} = \delta_{i1}, \ldots, y_{iT} = \delta_{iT} \mid X_i, u_i, \sum_{t=1}^{T} y_{it} = y_{i+}) = \frac{\exp(\sum\limits_{t=1}^{T} X_{it}\beta\delta_{it})}{\sum\limits_{d \in B_i} \exp(\sum\limits_{t=1}^{T} X_{it}\beta d_t)}$$

where

$$B_i = \left\{ d = (d_1, ..., d_T) \text{ such that } d_t \in \{0,1\} \text{and } \sum_{t=1}^{T} d_t = \sum_{t=1}^{T} y_{it} \right\}$$

The set B_i differs between individuals according to the value of $\sum\limits_{t=1}^{T} y_{it}$, i.e. the number of visits to state 1. Parameter β is estimated by maximising this conditional log-likelihood function. The estimator is consistent as $N \to \infty$, regardless of T (Andersen, 1970, Chamberlain, 1980, 1984, Hsiao, 1996). Nothing is known about its efficiency as in general conditional likelihood estimators are not efficient. Note that only the "moving" individuals are used in the computation of the conditional likelihood. Extensions of model (1.27) can be considered. For instance, Thomas (2005) develops the case where individual effect are multiplied by a time effect which is to be estimated.

The estimation of such a $T-$period model is also possible by reducing sequences of T observations into pairs of binary variables. Lee (2002) develop two interesting cases. First, the T periods can be chained sequentially two-by-two and a $T = 2$ conditional model can be estimated (as in Manski, 1987 see below). All pairs of periods two-by-two could also be considered. These decompositions will have an interest when generalizing conditional logit, when considering semi-parametric methods or more casually, as initial conditions for conditional maximum likelihood. It is why we now review the $T = 2$ case.

3.2.1 An example: the two period static logit model.

The conditional log-likelihood based on the logit model with T=2 computed with *moving* individuals is given by:

$$L = \sum_{d_i=1} \log \frac{\exp X_{i2}\beta}{\exp X_{i1}\beta + \exp X_{i2}\beta} + \sum_{d_i=0} \log \frac{\exp X_{i1}\beta}{\exp X_{i1}\beta + \exp X_{i2}\beta},$$

where for *moving* individuals, the binary variable d_i is:

$$\begin{cases} d_i = 1 & \text{if} \quad y_{i1} = 0, y_{i2} = 1 \\ d_i = 0 & \text{if} \quad y_{i1} = 1, y_{i2} = 0 \end{cases}$$

Denote $\Delta X_i = X_{i2} - X_{i1}$. The conditional log-likelihood becomes:

$$L = \sum_{i|d_i=1} \log \frac{\exp(\Delta X_i\beta)}{1 + \exp(\Delta X_i\beta)} + \sum_{i|d_i=0} \log \frac{1}{1 + \exp(\Delta X_i\beta)}$$

which is the expression of the log-likelihood of the usual logit model:

$$P(d_i = 1|\Delta X_i) = F(\Delta X_i\beta) \tag{1.28}$$

adjusted on the sub-sample of *moving* individuals. Note that the regressors do not include an intercept, since in the original model the intercept was absorbed by the individual effects.

3.2.2 A generalization. The consistency properties of conditional likelihood estimators are well known (Andersen, 1970) and lead to the interesting properties of conditional logit. This method has however been criticized on the ground that assuming a logistic function is a strong distributional assumption. When the errors v_{i1} and v_{i2} are independent, it can be shown that the conditional likelihood method is applicable only when the errors are logistic (Magnac (2004)). It is possible however to relax the independence assumption between errors v_{i1} and v_{i2} to develop a richer semi-parametric or parametric framework in the case of two periods. As above, pairing observations two-by-two presented by Lee (2002) can be used when the number of periods is larger.

The idea relies on writing the condition that the sum $y_{i1}+y_{i2}=1$ is a sufficient statistic in the sense that the following conditional probability does not depend on individual effects:

$$P(y_{i1}=1,y_{i2}=0 \mid X_i,u_i,\sum_{t=1}^{2} y_{it}=1)=P(y_{i1}=1,y_{i2}=0 \mid X_i,\sum_{t=1}^{2} y_{it}=1)$$

In that case, the development in the previous section can be repeated because the conditional likelihood function depends on parameter β and not on individual effects. It can be shown that we end up with an analog of equation (1.28) where distribution $F(.)$ is a general function which features and semi-parametric estimation are discussed in Magnac (2004).

3.3 Fixed effect Maximum Score

The methods discussed until section 3.2.2 are very attractive under one key condition, namely that the chosen distributional assumptions for the latent model are correct, otherwise the estimators will be typically inconsistent for the parameters of the model. However, since those functional restrictions are usually chosen for computational convenience instead of a priori plausibility, models that require less stringent assumptions or which are robust to violations of these assumptions, are attractive. Manski (1987) was the first to suggest a consistent estimator for fixed effect models in situations where the other approaches do not work. His estimator is a direct extension of the maximum score estimator for the binary model (Manski, 1975). The idea of this estimator for cross-sectional data is that if the *median* of the error term conditional on the regressors is zero, then observations with $X_i\beta > 0$ (resp. < 0) will have $P(y = 1 \,|X_i\beta > 0) > 0$ .5 (resp < 0.5). Under some regularity conditions this implies that $E\left\{sgn(2y_i - 1)sgn(X_i\beta)\right\}$ is uniquely maximised at the true value (in other words $(2y_i - 1)$and $(X_i\beta)$ should have the same sign). Therefore, the analogue estimator obtained by substituting expectations by means is consistent although not asymptotically normal and converges at a rate $N^{1/3}$ to a non-normal distribution (Kim and Pollard, 1990). There is however a smoothed version of this estimator where the sign function is substituted with a kernel type function, which is asymptotically normal and comes arbitrarily close to $\sqrt{N}$-convergence if tuning parameters are suitably chosen (Horowitz, 1992). However, Chamberlain (1992) shows that it is not possible of attaining a rate of $\sqrt{N}$ in the framework adopted by these papers.

Using a similar reasoning as in the conditional logit model and using the assumption that the distribution of the errors over time is stationary, Manski (1987) showed that, conditional on X:

$$P(y_2 = 1 \,|y_2 + y_1 = 1, X_i) > 0.5 \text{ if } (X_2 - X_1)\beta > 0$$

Therefore, for a given individual higher values of $X_t\beta$ are more likely to be associated with $y_t = 1$. In a similar fashion as the cross-sectional maximum score estimator, this suggests the following conditional maximum score estimator:

$$\hat{\beta}_N = \arg\max_{\beta} \sum_{i=1}^{N} sgn(y_{i2} - y_{i1})sgn[(X_{i2} - X_{i1})\beta]$$

For longer panels one can consider all possible pairs of observations over time:

$$\hat{\beta}_N = \arg\max_{\beta} \sum_{i=1}^{N} \sum_{s<t} sgn(y_{is} - y_{it})sgn[(X_{is} - X_{it})\beta]$$

The estimator has similar properties than the cross-sectional M-score estimator, in the sense that it is consistent under very weak conditions, but not asymptotically normal and converges at a rate slower than $\sqrt{N}$. Kyriazidou (1995) and Charlier, Melenberg, and van Soest (1995) show that the same 'smoothing trick' that worked for the cross-sectional M-score estimator also works for the conditional panel version. Hence, depending on the choice of smoothing parameters, the rate of convergence may come arbitrarily close to$\sqrt{N}$.

In practice, there are few applications of this estimator, since many difficulties arise : the solution of the optimisation problem is not unique, and the optimisation can be very complicated, because of the step function involved.

Other semi-parametric methods of estimation include Lee (1999) and Honoré and Lewbel (2002). In the first paper, an assumption about the dependence between individual effects and explanatory variables allows for the construction of method of moments estimator which is root-N consistent and asymptotically normal. In the second paper, another partial independence assumption is made as well as assumptions about the large support of one special continuous covariate. By linearizing the model (Lewbel, 2000), one can return to the reassuring world of linear models and difference out the individual effects. The reader is referred to the original papers in both cases.

3.4 GMM estimation

A possible solution to solving the problem posed by the presence of unobservable individual effect is to propose moment conditions which will be approximately satisfied provided that the individual effects are small, and estimators based on such moments (Laisney and Lechner, 2002). Consider the moment condition for any $t = 1, \ldots, T$:

$$E(y_t \,|X_i, u_i) = F(x_{it}\beta + u_i)$$

When the individual effect is close enough to the value of $\tilde{u}$, the first order Taylor approximation around $u = \tilde{u}$ is exact, so we can write for any $s, t = 1, \ldots, T$:

$$U - \tilde{u} = \frac{E[y_t \,|X_i, u_i] - F(X_t\beta + \tilde{u})}{f(X_t\beta + \tilde{u})} = \frac{E[y_s \,|X_i, u_i] - F(X_s\beta + \tilde{u})}{f(X_s\beta + \tilde{u})}$$

Thus, for any $s, t = 1, \ldots, T; s \neq t$, the following function,

$$m_{ts}(y, X; \beta) = \frac{y_t - F(X_t\beta - \tilde{u})}{f(X_t\beta - \tilde{u})} - \frac{y_s - F(X_s\beta - \tilde{u})}{f(X_s\beta - \tilde{u})}$$

has a conditional mean of zero at the true value of β, given $X = X_i$. It can be used as the basis for (almost) consistent estimation of the panel probit model with fixed effects close to $\tilde{u}$. Under standard regularity conditions, a GMM estimator of the coefficients for the time varying regressors of the panel model based on these moment functions is consistent (almost, given the Taylor approximation) and $\sqrt{N}$ asymptotically normal (Newey, 1993 ; Newey, McFadden, 1994).

3.5 Large-T Approximations

Finally, there are some new developments that are only briefly sketched here and that rely on large-T approximations in parametric binary models. The inspiration comes from Heckman's (1981a) pioneering work. Monte Carlo experiments can indeed be used to assess the magnitude of the bias of fixed effect estimators in binary probit or logit models as it was developed in the previous section. This bias due to the presence of incidental parameters is of order $O(T^{-1})$ in panel probit and for values around $T = 10$ the bias is found to be small (see also Greene, 2002).

A first direction for improving estimators is to assess and compute the bias either analytically or by using jackknife techniques as proposed by Hahn and Newey (2004). Under assumptions of independence over time of regressors and disturbances, bias-corrected estimators can be easily constructed. Hahn and Kuesteiner (2004) relax the assumption independence over time by proposing another analytical correction of the bias and that could also apply to the dynamic case (see next section).

The second direction relies on parameter orthogonalization. Inconsistency of fixed effect estimators occurs because the number of useful observations to estimate individual effects is fixed and equal to T and because there is contamination from the inconsistency of individual effect estimates into the parameters of interest. If, as in the Poisson count data example,[13] parameters of interest and individual effects can be factored out in the likelihood function (Lancaster, 2003) contamination is absent. Parameters are said to be orthogonal. These cases are not frequent however. The pionnering work of Cox and Reid (1987) uses a weaker notion of information orthogonality. At the true parameter values, the expectation of the cross derivative of the likelihood function w.r.t. the parameter of interest and the nuisance parameters is equal to zero. The invariance of likelihood methods to reparametrizations can then be used. The reparametrization which is interesting to use is the one (if it exists) that lead to information orthogonality. If this reparametrization is performed and if the nuisance parameters are integrated out in Bayesian settings, or concentrated out in classical settings, the bias of the ML estimator is of order $1/T^2$ instead of $1/T$ (in probability). For Probit (or other parametric) models, this method is proposed by Lancaster (2003) in a Bayesian setting. General theory in parametric non linear models in the Bayesian case is developed by Woutersen (2002). In the classical case, the panel static probit model is studied in a Monte Carlo experiment as an example by Arellano (2003) and in a dynamic case by Carro (2003). They show that for moderate T (4-6), the bias is small. It is smaller than the value for T advocated by

Heckman (1981a) though these values shall be theoretically validated in each instance where it is applied, as always when using Monte Carlo experiments about approximations.

4. Dynamic Models

In dynamic models where explanatory variables comprise lagged endogenous variables and other predetermined variables, we could further abandon the assumption that individual-and-period shocks and explanatory variables are independent. We distinguish again random and fixed effects models. This section is short not because the subject is unimportant but because the main ideas are extensions of the strict exogeneity case. There is one original issue however that we shall insist on, which is the choice of initial conditions.

4.1 Dynamic Random Effects Models

There are many potential sources of dynamics in econometric models. Some sources are easily dealt with in the framework of the last section: coefficients changing over time, lagged values of the strictly exogenous explanatory variables, correlation of random effects over time. There could also be true state dependence that is structural dependence on the lagged dependent variable or feedback effects of dependent variables on explanatory variables. Those explanatory variables are thus predetermined instead of strictly exogeneous. Most behavioral economic models using time-series or panel data are likely to be dynamic in this sense.

There are various dynamic discrete models as introduced by Heckman (1981a). The latent model that we study in this section, is written as:

$$y^*_{it} = \alpha y_{it-1} + X_{it}\beta + u_i + v_{it} \tag{1.29}$$

where individual effects u_i or individual-and-period specific effects v_{it} are or can be dependent of explanatory variables y_{it-1} and X_{it} and/or the future of these variables. It is in this sense that right-hand side variables are endogenous in this section. For simplicity we here consider one lag only and that v_{it} are independent of the past and present of (y_{it-1}, X_{it}).

As an alternative to this model (1.29), there is a class of models in which the lagged latent variable, y^*_{it-1}, is included among explanatory variables instead of the binary variable y_{it}. This type of dynamics is called habit persistence. Because recursive substitution techniques can be used – the lagged latent variable is replaced recursively by their expression (1.29) – these habit persistence models can be transformed into static models where explanatory variables include lags of the exogenous variables and where some care should be taken with the initial condition, y^*_{i1}. These types of models are discussed briefly in Heckman (1981). Estimation of the structural parameters in the case of binary choice is detailed in Lechner (1993). Moreover, this framework does not accomodate weak endogeneity which is one of the focus of this section.

4.1.1 Initial conditions. When the lagged endogenous variable is present, there is an initial condition problem as in the linear case though it is more diffcult to deal with. Assuming for the moment that there are no other explanatory variables, $\beta = 0$, the likelihood function is written by conditioning on individual effects as in the previous section:

$$l(y_{iT}, ., y_{i2}, y_{i1} \mid u_i) = \prod_{t=2}^{T} l(y_{it} \mid y_{it-1}, u_i) l(y_{i1} \mid u_i)$$

It is obvious that one needs additional information for deriving $l(y_{i1} \mid u_i)$ that model (1.29) is not providing. It is analogous to the linear case and the assumptions that initial conditions are exogenous or that initial conditions are obtained by initializing the process in the infinite past were soon seen to be too strong or misplaced. They are generally strongly rejected by the data. Heckman (1981) proposed to use an auxiliary assumption such as:

$$y^*_{i1} = \theta u_i + v^0_{i1}. \tag{1.30}$$

The complete likelihood function is then obtained by integrating out, u_i, as before.

Another route was suggested by Wooldridge (2000) or Arellano & Carrasco, (2002). Instead of using the complete joint likelihood function, they resort to the following conditional likelihood function:

$$l(y_{iT}, ., y_{i2} \mid y_{i1}, u_i) = \prod_{t=2}^{T} l(y_{it} \mid y_{it-1}, u_i).$$

When integrating out u_i, one now needs to choose the conditional distribution function $f(u_i \mid y_1)$ which might be written as the auxiliary model which marries well with the approach of Chamberlain seen above:

$$u_i = \theta y_{i1} + \eta_i \tag{1.31}$$

It should be noted that one loses information and that it is not immediately clear whether restriction (1.30) is more restrictive than (1.31) in particular when other explanatory variables are present in the model.

4.1.2 Monte Carlo Experiments of Simulation Methods. In the literature, some papers report Monte Carlo experiments of random effects dynamic models estimated by simulation (Keane, 1994, Chib et Jeliazkov, 2002, Lee, 1997). There seems to be a consensus on a few results. Estimates of the autoregressive parameter seem to be downward biased while parameters of the variance of random effects can be upward or downward biased according to the model (Lee, 1997). Biases increase when serial correlation is stronger though it can be counteracted by increasing the number of draws either for SML or MSM as well Gibbs sampling. Biases also increase when the number of periods increases. Misspecification of initial conditions introduces fairly large biases in the estimation.

4.1.3 A Projection Method. For treating the weakly endogenous case, there has been an interesting suggestion proposed by Arellano & Carrasco (2002). Let $\omega_{it} = (y_{it-1}, X_{it})$ be the relevant conditional information in period t that is grouped into the information set, $\omega_i^t = (\omega_{it}, \omega_i^{t-1})$ where ω_i^0 is the empty set. Variables ω_i^t summarize the relevant past of the process until period t, that is the sequence of lagged endogeneous variables, explanatory variables and their lags and any other piece of information such as instruments for instance. Assume that $\varepsilon_{it} = u_i + v_{it}$ is such that:

$$\varepsilon_{it} \mid \omega_i^t \rightsquigarrow N(E(u_i \mid \omega_i^t), \sigma_t^2)$$

where independence between v_{it} and the information set ω_i^t has been used. Thus, it rules out serial correlation in the usual sense[14] while allowing for feedback. It thus constitutes a generalization of the setting of the projection method of Chamberlain (1980) and Newey (1994) that we presented in the previous section.

The sequence of conditional means $E(u_i \mid \omega_i^t)$ are related by the moment conditions:

$$E(E(u_i \mid \omega_i^t) \mid \omega_i^{t-1}) = E(u_i \mid \omega_i^{t-1}) \tag{1.32}$$

Write now the conditional means:

$$E(y_{it} \mid \omega_i^t) = \Phi(\frac{\alpha y_{it-1} + X_{it}\beta + E(u_i \mid \omega_i^t)}{\sigma_t})$$

which translates into:

$$\sigma_t.\Phi^{-1}(E(y_{it} \mid \omega_i^t)) = \alpha y_{it-1} + X_{it}\beta + E(u_i \mid \omega_i^t)$$

The moment condition (1.32) is thus:

$$\begin{aligned} &E(\sigma_t.\Phi^{-1}(E(y_{it} \mid \omega_i^t)) - (\alpha y_{it-1} + X_{it}\beta) \mid \omega_i^{t-1}) = \\ &\qquad = \sigma_{t-1}.\Phi^{-1}(E(y_{it} \mid \omega_i^{t-1})) - (\alpha y_{it-2} + X_{it-1}\beta) \end{aligned}$$

As before, some nonparametric estimates of $E(y_{it} \mid \omega_i^t)$ can be obtained and plugged in this moment condition.

As it is formally identifical to the approach proposed by Newey (1994), the same remarks can be addressed to this approach. There may however be a curse of dimensionality coming in because the dimension of ω_i^t is growing with the number of periods. Arellano & Carrasco (2002) proposes simplifications and the reader is referred to the original paper.

4.2 Dynamic Fixed Effects Models

Chamberlain (1985) extends the conditional logit method to the case where the lagged endogenous variable is the only covariate (see also Magnac, 2000, for multinomial and dynamic models where lags can be larger than 1). Sufficient statistics are now a vector of three variables. On top of the sum of binary variables, the binary variables at the first and last period are added to the list. For instance, in the case where only one lag is used, the smallest number of periods for identification is equal to 4 and the useful information is contained in the intermediate periods from $t = 2$ to $T - 1$. The main drawback of this method is that, in the logit case and in the model with one lag, the sum of binary variables, the first and last values of the binary variables are not sufficient statistics if other explanatory variables are present in the model.

If explanatory variables are discrete, the idea proposed by Honoré & Kyriazidou (2000) is to consider only the observations such that explanatory variables are constant in the intermediate periods from $t = 2$ to $T - 1$. Conditional to the values of these explanatory variables, the sum of binary variables, the first and last values of the binary variables are now sufficient statistics. In order to accomodate continuous variables, Honoré & Kyriazidou (2000) proposes to use observations such that explanatory variables are approximately constant in the intermediate periods from 2 to $T - 1$. The statistics described above are approximately sufficient. Observations can be weighted according to the degree of such an approximation. Under some conditions the estimator is consistent and asymptotically normal, but due to the nonparametric part, its convergence rate is less than $\sqrt{N}$. Note also that this construction rules out time dummies, which cannot by definition be similar in two periods.

Notes

1. One needs to assume independence between errors and regressors instead of assuming that correlations are equal to zero because of the non-linearity of the conditional expectation of the dependent variable **with respect to individual effects**.

2. See for example the discussion in Chamberlain, 1984

[3] As it can be found in STATA for instance.

[4] Hajivassiliou and McFadden (1998) first propose to simulate the numerator and the denominator separately. Of course, this method does not lead to unbiased simulation because the ratio is not linear but, still, as simulators are asymptotically unbiased, those MSS estimators are consistent whenever H tends to infinity. The authors furthermore argue that using the same random draws for the denominator and the numerator decreases the noise. The other method based on Gobbs resampling seems expensive in terms of computations using large samples though it is asymptocally unbiased as soon as H tends to infinity faster than $log(N)$.

[5] See Chib and Greenberg (1998) to assess how to do the division into blocks according to the identifying or other restrictions on parameter β or on matrix Ω

[6] The following section heavily draws from Bertschek and Lechner (1998)

[7] For all details, the reader is referred to Bertschek and Lechner, 1998

[8] Lechner (1995) proposes specification tests for this estimator.

[9] The so-called Mundlak (1978) approach is even more specific since individual effects u_i are written as a function of averages of covariates, $\frac{1}{T}\sum_{t=1}^{T} x_{it}$ only and a redefined individual effect η_i.

[10] Not in all cases, the example of count data being prominent (Lancaster, 1998).

[11] It is however possible to define restrictions to identify these effects, see Chapter XXX

[12] The claim that a parametric distributional assumption of individual effects is needed for the identification of causal treatment effects is however overly strong. What is true is that the estimation of the conditional distribution function of individual effects is almost never considered though it can be under much weaker assumptions than parametric ones.

[13] As described in Montalvo (1997) and Blundell, Griffith and Windmeijer (2002).

[14] Individual-and-period v_{it-1} is not included in ω_i^t, only y_{it-1} is.

5. References

Altonji J.G., and L.M., Segal, 1996, “Small-sample bias in GMM estimation of covariance structures”, *Journal of Business Economics and Statistics*, 14: 353-366.

Andersen, E.B., 1970: “Asymptotic Properties of Conditional Maximum Likelihood Estimators”, *Journal of the Royal Statistic Society*, Series B, 32, 283-301.

Angrist, J.D., 2001, “Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice”, *Journal of Business Economics and Statistics,* 19:2-16.

Arellano M., 2003, “Discrete Choice with Panel Data”, forthcoming *Investigaciones Economicas.*

Arellano M. and R. Carrasco, 2003, “Binary Choice Panel Data Models with Predetermined Variables”, *Journal of Econometrics*, 115, 125-157.

Arellano M. and B. Honoré, 2001, “Panel Data Models: Some Recent Developments”, in eds J. Heckman and E. Leamer, *Handbook of Econometrics,* V(53):3229-3296.

Avery, R.B, L.P., Hansen and V.J., Hotz, 1983, “Multiperiod Probit models and orthogonality condition estimation”, *International Economic Review*, 24:21-35.

Baltagi, B.H., 2000, *Econometric Analysis of Panel Data,* Wiley: London.

Butler J. and R.Moffitt, 1982, " A computationally efficient quadrature procedure for the one-factor multinomial probit model ", *Econometrica*, Vol 50, N ° 3, 761-764.

Bertschek I. and M. Lechner, 1998, "Convenient estimators for the panel probit model", *Journal of Econometrics*, 87, 329-371.

Breitung, J. and M. Lechner, 1997, "Some GMM Estimation Methods and Specification Tests for Nonlinear Models", in L.Mátyás, P.Sevestre (Eds.), *The Econometrics of Panel Data*, 2^{nd}ed., Dordrecht: Kluwer, 583-612, 1996.

Carro J.M., 2003, "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", Working paper, CEMFI, 0304.

Chamberlain G.,1980, "Analysis of covariance with qualitative data", *Review of Economic Studies*, 47, 225-238.

Chamberlain G., 1984, : " Panel Data ", in Z. Griliches and M.D. Intrilligator ed., *Handbook of Econometrics*, vol II, ch 22, Elsevier Science, pp 1248-1318.

Chamberlain, G., 1985, "Heterogeneity, Omitted Variable Bias and Duration Dependence", in *Longitudinal Analysis of Labor Market Data,* in eds J.J. Heckman and B. Singer, Cambridge UP: Cambridge.

Chamberlain, G., 1992, "Binary Response Models for Panel Data: Identification and Information", *mimeo*, Harvard University.

Charlier, E., B. Melenberg, and A. van Soest, 1995, "A Smoothed Maximum Score Estimator for the Binary Choice Panel Model and an Application to Labour Force Participation", *Statistica Neerlandica*, 49, 324-342.

Chen, S., 1998, "Root-N consistent estimation of a panel data sample selection model", unpublished manuscript, Hong Kong university.

Chib, S., 2001, "Markov Chain Monte Carlo Methods: Computation and Inference", in eds J. Heckman and E. Leamer, *Handbook of Econometrics,* V(57):3570-3649.

Chib, S., and E. Greenberg, 1998, "Analysis of Multivariate Probit Models", *Biometrika,* 85:347-61.

Chib, S., and I. Jeliazkov, 2002, "Semiparametric Hierarchical Bayes Analysis of Discrete Panel Data with State Dependence", Washington University, working paper.

Cox, D.R., and M., Reid, 1987, "Parameter Orthogonality and Approximate Conditional Inference", *Journal of the Royal Statistical Society,* Series B, 49:1-39.

Crépon B. and J. Mairesse, 1996, "The Chamberlain Approach to Panel Data: An Overview and Some Simulation Experiments", in L. Matyas and P. Sevestre eds, *The Econometrics of Panel Data*, Kluwer:Amsterdam.

Geweke J., M. Keane and D.E. Runkle, 1997, " Statistical inference in the multinomial multiperiod probit model ", *Journal of Econometrics*, 80, 125-165.

Geweke J.F., M., Keane, 2001, "Computationally Intensive Methods for Integration inEconometrics", in eds J. Heckman and E. Leamer, *Handbook of Econometrics,* V(56):3465-3568.

Gouriéroux C. and A.Monfort, 1993 : " Simulation-based inference : A survey with special reference to panel data models ", *Journal of Econometrics*, 59, 5-33.

Gouriéroux C. and A., Monfort, 1996 : *Simulation-based Econometric Methods*, Louvain: CORE Lecture Series.

Gouriéroux C., A., Monfort and A.Trognon, 1984, "Pseudo-likelihood methods - Theory", *Econometrica,* 52: 681-700.

Greene, W., 2002, "The Bias of the Fixed Effects Estimator in Non Linear Models", New York University, unpublished manuscript.

Greene, W., 2003, *Econometric Analysis,* 5th edition, Prentice Hall: Englewood Cliffs.

Guilkey, D.K. and Murphy, J.L., 1993, "Estimation and Testing in the Random Effects Probit Model", *Journal of Econometrics*, 59, 301-317.

Hahn J. and G. Kuersteiner, 2004, "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects", unpublished manuscript.

Hahn J. and W. Newey, 2004, "Jackknife and Analytical Bias Reduction for Nonlinear Panel Data Models", *Econometrica,* 72:1295-1319.

Hajivassiliou V. and D., Mc Fadden, 1998, " The Method of Simulated Scores for the Estimation of LDV Models ", *Econometrica*, vol 66, 863-896.

Hajivassiliou V., Mc Fadden D. and P. Ruud, 1996, " Simulation of multivariate normal rectangle probabilities and their derivatives. Theorical and computational results ", *Journal of Econometrics*, 72, 85-134.

Heckman, J.J., 1981a, : "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time - Discrete Data Stochastic Process and Some Monte-Carlo Evidence," in C. Manski and D. McFadden, Ed., *Structural Analysis of Discrete Data*, MIT Press, Cambridge, MA, 179-195.

Heckman, J.J., 1981b, : “Statistical Models for Discrete Panel Data” in C. Manski and D. McFadden, Ed., *Structural Analysis of Discrete Data*, MIT Press, Cambridge, MA, 114:178.

Heckman, J.J. and B., Singer, 1984, A method for minimizing the impact of distributional assumptions in econometric models for duration data”, *Econometrica*, 52:271-320.

Honoré, B., 2002, “Non-Linear Models with Panel Data”, WP CEMMAP, 13/02.

Honoré, B., and E., Kyriazidou, 2000, “Panel Data Discrete Choice Modles with Lagged Dependent Variables”, *Econometrica,* 68:839-74.

Honoré, B.E., and A., Lewbel, 2002, “Semiparametric Binary Choice Panel Data Models without Strict Exogeneity”, *Econometrica*, 70:2053-2063.

Horowitz, J., 1992), ”A Smoothed Maximum Score Estimator for the Binary Response Model”, *Econometrica*, 60, 505-531.

Hsiao, C., 1986, *Panel Data,* Cambridge University Press.

Hsiao, C., 1992, “Logit and Probit Models”, in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data : Handbook of Theory and Applications*, chapter 11: 223-241, Kluwer: Amsterdam.

Hsiao, C., 1996, “Logit and Probit Models”, in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data : Handbook of Theory and Applications*, 2nd edition, chapter 16: 410-428 , Kluwer: Amsterdam.

Inkman, J., 2000, “ Misspecified heteroskedasticity in the panel probit model: A small sample comparison of GMM and SML estimators”, *Journal of Econometrics*, 97: 227-259.

Kamionka, T., 1998, “Simulated maximum likelihood estimation in transition models, *Econometrics Journal,* 1:C12*9-153.*

Keane, M.P., 1994, “A Computationally Efficient Practical Simulation Estimator for Panel Data”, *Econometrica*, 62:95-116.

Kim, J. and D. Pollard (1990) : “Cube Root Asymptotics”, *Annals of Statistics*, 18, 191-219.

Kyriazidou, E., 1995, *Essays in Estimation and Testing of Econometric Models*, Ph.D. dissertation, Northwestern University.

Laisney F. and M. Lechner, 2002,: “Almost Consistent Estimation of Panel Probit Models with ’Small’ Fixed Effects”, *Discussion paper* no. 2002-15, University of St. Gallen.

Lancaster, A., 2000, “The Incidental Parameter Problem since 1948”, *Journal of Econometrics,* 95:391-413.

Lancaster, **A**., 2003, *An Introduction to Modern Bayesian Econometrics,*c frotcoming Blackwell:Oxford.

Lechner, M. 1993: "Estimation of Limited Dependent Variable Habit Persistence Models on Panel Data with an Application to the Dynamics of Self-employment in the Former East Germany", in Bunzel, H., Jensen, P. and Westergård-Nielson, N. (eds.), *Panel Data and Labour Market Dynamics*, Amsterdam: North-Holland, 263-283.

Lechner, M., 1995, "Some Specification Tests for Probit Models Estimated on Panel Data", *Journal of Business & Economic Statistics*, 13, 475-488, 1995.

Lee, L.F., 1992, "On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models", *Econometric Theory, 8:518-552.*

Lee, L.F., 1995, "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models", *Econometric Theory, 11:437-83.*

Lee, L.F., 1997, "Simulated Maximum Likelihood Estimation of Dynamic Discrete Choice Statistical Models: Some Monte carlo Results", *Journal of Econometrics*, 82:1-35.

Lee, L.F., 2000, "A numerically stable quadrature procedure for the one-factor random component discrete choice model", *Journal of Econometrics*, 95, 117-129.

Lee, M.J., 1999, "A Root-n Consistent Semiparametric Estimator for Related-Effect Binary Response Panel Data", *Econometrica*, 67:427-33.

Lee M.J., 2002, *Panel Data Econometrics,* Academic Press: New York.

Lewbel, A., 2000, "Semiparametric Qualitative Response Model Estimation with Unknown Heteroskedasticity or Instrumental Variables", *Journal of Econometrics*, 97:145-77.

McCulloch, R. and P. E.,Rossi, 1994, " An exact likelihood analysis of the multinomial probit model", *Journal of Econometrics*, 64, 207-240.

McFadden D., 1989, " A Method of Simulated Moments for Estimation od Discrete Response Models without Numerical Integration ", *Econometrica*, vol 57, pp 995-1026.

Magnac, T., 2000, "State dependence and unobserved heterogeneity in youth employment histories", *The Economic Journal,* 110:805-837.

Magnac, T. 2004, "Binary Variables and Sufficiency: Generalizing the Conditional Logit", *Econometrica,* 72:1859-1876.

Manski, C.F. 1975, "Maximum Score Estimation of the Stochastic Utility Model", *Journal of Econometrics*, 3, 205-228.

Manski, C.F., 1987, "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, 357-362.

Manski, C.F., 1988, "Identification of Binary Response Models", *Journal of the American Statistical Association,* 83:729-738.

Mroz T., 1999, " Discrete factor approximation in simultaneous equation models : estimating the impact of a dummy endogenous variable on continuous outcome ", *Journal of Econometrics*, vol 92, 233-274.

Newey, W., 1993, "Efficient estimation of models with conditional moment restrictions", in Maddala, G.S., Rao, C., Vinod, H. (Eds.), *Handbook of Statistics*, Vol. 11, Ch. 16, North-Holland, Amsterdam.

Newey, W., 1994, "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, 62:1349-82.

Newey, W.K. and McFadden, D., 1994, "Large Sample Estimation and Hypothesis Testing", in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol 4, 2113-2245, Amsterdam: North-Holland.

Pagan A. and A. Ullah, 1998, *Nonparametric Econometrics,* Cambridge UP, Cambridge.

Robinson P.M., 1982, "On the Asymptotic Properties of Estimators of Models containing Limited Dependent Variables", *Econometrica*, 50:27-41.

Sevestre P., 2002, *Économétrie des données de panel*, Dunod: Paris.

Thomas, A., 2003, "Consistent Estimation of Binary-choice Panel Data Models with Heterogeneous Linear Trends", LEERNA-INRA Toulouse, unpublished manuscript.

Train, K., 2002, *Discrete Choices with Simulation,* Cambridge U.P:Cambridge.

Wooldridge J., 2000, *Introductory Econometrics*, 2nd edition, South-Western College Publishing.

Wooldridge J., 2002, "Simple Solutions to the Initial Conditions Problem in Dynamic Non Linear Panel Data Models with Unobserved Heterogeneity", WP CEMMAP, London, 18/02.

Woutersen T., 2002, "Robustness against Incidental Parameters", Western Ontario, unpublished manuscript.